PERSONALITY PROCESSES AND INDIVIDUAL DIFFERENCES

# Evaluating Replicability of Factors in the Revised NEO Personality Inventory: Confirmatory Factor Analysis Versus Procrustes Rotation

Robert R. McCrae, Alan B. Zonderman,
and Paul T. Costa, Jr.
Gerontology Research Center, National Institute on Aging,
National Institutes of Health, Baltimore, Maryland

Michael H. Bond
Chinese University of Hong Kong

Sampo V. Paunonen
University of Western Ontario

Despite the empirical robustness of the 5-factor model of personality, recent confirmatory factor analyses (CFAs) of NEO Personality Inventory (NEO-PI) data suggest they do not fit the hypothesized model. In a replication study of 229 adults, a series of CFAs showed that Revised NEO-PI scales are not simple-structured but do approximate the normative 5-factor structure. CFA goodness-of-fit indices, however, were not high. Comparability analyses showed that no more than 5 factors were replicable, which calls into question some assumptions underlying the use of CFA. An alternative method that uses targeted rotation was presented and illustrated with data from Chinese and Japanese versions of the Revised NEO-PI that clearly replicated the 5-factor structure.

If an aeronautical engineer announced that the latest supercomputer simulation proved that monoplanes cannot fly, we would not rush to ground the airfleets of the world. We would instead conclude that the computer simulation was fatally flawed. It is the essence of empiricism that conceptual models, no matter how mathematically elegant, are abandoned when they fail to lead to accurate predictions of known facts. In this article we argue that maximum likelihood confirmatory factor analysis (CFA), as it has typically been applied in investigating personality structure, is systematically flawed: Its statistical indices reject models that are empirically replicable (see Study 1) and accept models that are not (see Study 2). We also propose an alternative approach to the statistical evaluation of factor replicability that yields more reasonable results.

## The Five-Factor Model and Its Statistical Evaluation

Over the past decade, studies of natural language adjectives (Goldberg, 1990; Ostendorf, 1990), California Q-Set (Block, 196 1) items (Lanning, 1994; McCrae, Costa, & Busch, 1986), and personality questionnaires (McCrae, 1989; Paunonen, Jackson, Trzebinski, & Forsterling, 1992) have converged on a representation of personality trait structure in terms of five orthogonal factors-the Big Five, or five-factor model (FFM). The Revised NEO Personality Inventory (NEO-PI-R, Costa & McCrae, 1992c) was designed to operationalize the FFM by measuring six specific traits, or facets, that define each of the five factors: Neuroticism (N), Extraversion (E), Openness to Experience (0), Agreeableness (A), and Conscientiousness (C).

Exploratory factor analyses that have used varimax rotation of principal components in younger and older, White and non-White, and male and female subsamples have shown very similar five-factor structures that supported the theoretical model (Costa, McCrae, & Dye, 199 1). Furthermore, that structure has been essentially replicated in self-reports from independent samples of adults (Costa & McCrae, 1992a; Piedmont & Weinstein, 1993) and college students (Costa & McCrae, 1994), in observer ratings from peers (Costa & McCrae, 1992d) and from spouses ( McCrae,

Robert R. McCrae, Alan B. Zonderman, and Paul T. Costa, Jr., Laboratory of Personality and Cognition, Gerontology Research Center, National Institute on Aging, National Institutes of Health, Baltimore, Maryland; Michael H. Bond, Department of Psychology, Chinese University of Hong Kong; Sampo V. Paunonen, Department of Psychology, University of Western Ontario, London, Ontario, Canada.

This research was supported in part by Social Sciences and Humanities Research Council of Canada Research Grant 410-91-1282. Portions of this article were presented at the 102nd Annual Convention of the American Psychological Association, August 1994, Los Angeles. We thank Yoshiko Shimonaka for allowing us to report analyses of the Japanese NEO-PI-R; Ross R. Vickers, Jr., Robert Cudeck, and A. Timothy Church for consultation on LISREL analyses; Ralph L. Piedmont for supplying the SPSS version of the program in the Appendix; and Carmi Schooler, Lewis R. Goldberg, John M. Digman, Gerard Saucier, Clarence McCormick, Peter Borkenau, Stephen G. West, Kevin B. Lanning, and Marco Perugini for helpful comments.

Correspondence concerning this article should be addressed to Robert R. McCrae, Personality, Stress, and Coping Section, Gerontology Research Center, 4940 Eastern Avenue, Baltimore, Maryland 2 1224. Electronic mail may be sent via the Intemet to jeffm@mvx.grc.nia.nih.gov.

Journal of Personality and Social Psychology, 1996, Vol. 70, No. 3, 552-566

1994a), and in versions of the instrument translated into German (Borkenau & Ostendorf, 1990) and Hebrew (Montag & Levin, 1994). The FFM, and in particular its operationalization in the NEO-PI-R, is clearly a model that "flies."

Recent CFAs, however, seem to have cast doubt on the structure of the NEO-PI-R. Borkenau and Ostendorf (1990), in the first CFA of the FFM, found that "even the least restrictive model ... did not appropriately account for the data" (p. 522). Church and Burke (1994), in a detailed analysis of the (unrevised) NEO-PI in a student sample, found that a simple structure model did not fit well and that a more complex model, based on results from an adult sample (McCrae & Costa, 1989a), "suggested at best a fair fit" (Church & Burke, 1994, p. 105). A reasonably good fit was obtained by Church and Burke in cross-validation of a model generated in the student data, but only when a number of ad hoc factor loadings and residual covariances were added to the model. Parker, Bagby, and Summerfeldt ( 1993) analyzed the intercorrelations among the NEO-PI-R facet scales in the normative data (Costa & McCrae, 1992c). They considered models in which facets were assigned to a single factor as well as somewhat more complex models in which secondary loadings suggested by previous studies were also included. They reported that "none of the models were [sic] found to be a satisfactory approximation of the NEO-PI-R data" (p. 464) and that there were substantial correlations among the factors. Panter, Tanaka, and Hoyle ( 1994) also concluded from an interbattery CFA that "the five factors are by no means orthogonal" (p. 134).

Why are CFA results at such variance with other empirical evidence of robust replicability of the NEO-PI-R structure? There are two possible reasons: The CFA technique may be inappropriately applied, or there may be fundamental problems with the CFA method itself.

### Simple Structure in Personality Measures

The question of the applicability of CFA to NEO-PI-R data has been raised by Church and Burke (1994), who claimed that CFA techniques are best suited to the analysis of simple structure models. To preserve the integrity of its statistical tests, CFA users are generally cautioned to test only clearly specified models and to alter the model only when there is a good theoretical rationale for the change. One of the strengths of CFA is its flexibility, and the technique can in principle be applied to extremely complex models suggested by prior results (as we show in Study 1), but hitherto most researchers have not considered the replication of a complete factor matrix to be a theoretically justifiable hypothesis.

In typical applications, therefore, variables are hypothesized to load on a single specified factor, and their loadings on other factors are fixed at zero. Maximum likelihood estimates are then used to determine the optimal values for the hypothesized factor loadings, and the fit of this solution (i.e., the adequacy with which the original correlation matrix can be reproduced from the factor loadings) is then evaluated. In real data the secondary loadings of variables are rarely exactly zero, but in simple structure models they are assumed to fluctuate randomly about zero. If, however, small loadings are in fact meaningful, CFA with a simple structure model may not fit well.

The factors of the FFM were initially identified with rotation procedures (like varimax) designed to approximate simple structure. Yet the FFM does not postulate perfect simple structure-that is, it does not assume that all personality traits define one and only one factor. Decades of research on the interpersonal circumplex (e.g., Wiggins, 1979) have shown that many important interpersonal traits fall between the orthogonal axes of E and A, and De Raad, Hendriks, and Hofstee ( 1992) extended this observation to other pairs of the five factors. There is no theoretical reason why traits should not have meaningful loadings on three, four, or five factors.

The NEO-PI-R appears to adopt a strict simple structure model, because (as a conceptual and scoring convenience) each of its 30 facet scales is assigned to a single domain, and domain scales-the sum of the six assigned facets-are used to estimate the five factors. In fact, however, several facets have large secondary loadings that are both meaningful and replicable. For example, Angry Hostility is considered a facet of N and typically has its largest loading on that factor. People who are low in A are also prone to experience anger, however, so Angry Hostility also has a large negative loading on the A factor (Church & Burke, 1994; Costa & McCrae, 1992c). In recognition of that fact, the authors of the NEO-PI-R advocate factor scores calculated from all 30 facets as the preferred measure of the FFM factors (Costa & McCrae, 1992c), and these factor scores are routinely calculated and used in the NEO-PI-R computer interpretive report.

It is possible to make allowances for secondary loadings in CFA by fixing the loadings at a priori values other than zero. When Parker et al. ( 1993) included salient (±.40) secondary loadings from previous studies in their analyses, they improved the fit of the model, as did Church and Burke (1994), who regarded loadings as small as ±.20 as salient secondaries. One hypothesis suggested by these considerations is that CFA fit should be improved by increasing the number of secondary loadings specified. This approach suggests that the most appropriate NEO-PI-R model to test is not the simple structure organization of facets into domains but the full 30 X 5 matrix of factor loadings reported for the NEO-PI-R's normative sample.

### Obliquity Versus Orthogonality

A second, related problem with the application of CFA to the NEO-PI-R concerns the orthogonality of the factors. Although the factors of the FFM are conceptualized as being orthogonal (Costa & McCrae, 1995b; Goldberg, 1993), and uncorrelated scales measuring them might in principle be constructed, NEO-PI-R domain scores consistently show nontrivial intercorrelations. These associations are due chiefly to the selection of facets to represent each domain (Costa & McCrae, 1992b). For example, the N domain includes Impulsiveness and Vulnerability facet scales that are negatively related to C, but it does not happen to include an Obsessiveness scale that would show a positive relation to C (Costa & McCrae, 1995a). As a result, there is a substantial negative correlation between NEO-PI-R N and C domain scales.

When, in oblique CFA, secondary loadings are fixed at zero, factors can be based only on the facets assigned to domains, and correlations among factors will tend to mirror the correlations among domains. These simple structure models of NEO-PI-R data will be oblique, sometimes strongly so, leading some researchers-like Parker et al. (1993)—to "challenge the view that the NEO-PI-R assesses five distinct personality dimensions" (p. 465). As Church and Burke (1994) argued, however, orthogonahty can be preserved if strict simple structure is abandoned. Factor scores based on weighted combinations of all 30 facet scales may provide uncorrelated measures.

These considerations lead to the following hypothesis: In CFA analyses of NEO-PI-R data, oblique factors will be clearly superior to orthogonal factors only when simple structure models are tested. When secondary loadings are specified a priori on the basis of the scales' normative structure, orthogonal factors will fit as well as oblique factors. In Study 1 we tested the hypotheses that fit would improve and that obliquity of the factors would be reduced by increasing the number of secondary loadings specified.

### *CFA and Its Alternatives*

Even proponents of CFA acknowledge a long list of problems with the technique, ranging from technical difficulties in estimation of some models to the cost in time and effort involved (e.g., Church & Burke, 1994; Jackson & Chan, 1980; Velicer & Jackson, 1990). The major advantage claimed for CFA is its ability to provide statistical tests of the fit of empirical data to different theoretical models. Yet it has been known for years that the chi-square test on which most measures of fit are based is problematic (Bentler & Bonett, 1980; Hu, Bentler, & Kano, 1992; Marsh, Balla, & McDonald, 1988). The statistic is directly related to sample size, and virtually any model will be rejected if based on a sufficiently large sample. Indeed, researchers using the chi-square test are penalized for analyzing large samples. A variety of alternative measures of goodness-of-fit have been suggested, but their interpretation and relative merits are not yet clear, and they do not yield tests of statistical significance.

CFA is closely akin to exploratory maximum likelihood factor analysis (EMLFA), which incorporates a version of the chi-square statistic to determine the number of factors needed to obtain a good fit. In practice, it is recognized that this chi-square test often leads to overextraction of factors (Gorsuch, 1983). That judgment is usually based on an examination of EMLFA results that typically include small, uninterpretable factors. More formally, overextraction might be defined as extracting more factors than are replicable in comparable independent samples of a sufficient size. Data showing that chi-square tests lead to overextraction in this sense call into question the appropriateness of those tests in both exploratory and confirmatory maximum likelihood models (Borkenau & Ostendorf, 1990). In Study 2 we examined the rephcability of NEO-PI-R factors suggested by chi-square tests.

Although there are problems with the CFA approach, the need for some formal evaluation of factor rephcability remains. The usual practice of comparing varimax factors from independent exploratory factor analyses provides powerful evidence of replicability when the same factors are found, but is hard to interpret when different factors are found. One promising alternative involves some form of Procrustes rotation (Barrett, 1986). Procrustes rotations, however, seem to be widely distrusted (e.g., Gorsuch, 1983); in Study 3 we provide arguments for the legitimacy of the procedure and report data showing that it yields meaningful conclusions about the factor structure of the NEO-PI-R.

## Study 1: Confirmatory Factor Analyses of the NEO-PI-R in a Community Sample

As part of a study of alternative models of personality structure, the NEO-PI-R was administered to a sample of 96 men and 133 women on a waiting list to join the Baltimore Longitudinal Study of Aging (BLSA; Shock et al., 1984). Like most BLSA participants, these individuals were generally healthy and well-educated; they ranged in age from 26 to 87 years (see Costa & McCrae, 1995a, for a more detailed description). Although slightly lower in N (combined-sex $T = 47$) and slightly higher in 0 ($T = 54$), this waiting list sample appears to be comparable to the NEO-PI-R adult normative sample, which included BLSA participants (Costa & McCrae, 1992c).

We examined the intercorrelations among the 30 NEO-PI-R facets in a series of CFA models, using LISREL 7 (Jöreskog & Sörbom, 1988). In each case, five factors were examined. We calculated a null *mode/* (Model 1), which assumes no common factors, to provide a baseline for evaluating the fit of other models. Four orthogonal models (Models 2a–2d) were then considered in which 30 free parameters (six facets for each of five factors) were estimated. In each model, the analysis determined the optimal loading of each facet scale on the factor to which it was primarily assigned (e.g., the loadings for Anxiety, Angry Hostility, Depression, Self-Consciousness, Impulsiveness, and Vulnerability on the N factor). The four models differed in the a priori values fixed for the remaining 120 factor loadings. We used values taken from a varimax-rotated principal-components analysis of the normative sample data (Costa & McCrae, 1992c; see Appendix for values) as the source of hypothesized a priori values.

In a *simple structure mode/* (Model 2a), all 120 fixed loadings were set to zero. In a *salient-loadings model* (Model 2b), five secondary loadings greater than ±.40 in the normative sample were fixed at the normative value; the other fixed loadings were set to zero. In a *modest-loadings model* (Model 2c), 32 secondary loadings greater than ±.20 in the normative sample were fixed at that value; the other loadings were set to zero. In a *complete model* (Model 2d), all 120 secondary loadings were fixed at their corresponding value in the normative sample. If, as hypothesized, the full factor matrix of the normative data provides the best model of the structure of the NEO-PI-R, these four models should show increasing fit.

We then re-estimated Models 2a–2d as Models 3a–3d by relaxing the constraint that the factors be orthogonal. In general, oblique factors will show a better fit than will orthogonal factors; we hypothesized, however, that the factor intercorrelations and the improvement in fit of each oblique model over the corresponding orthogonal model would diminish from the oblique simple structure model (3a) to the oblique complete model (3d).

Models 2b–2d and 3b–3d employed fixed values taken from the normative matrix of varimax-rotated principal components. CFA analyses, however, estimate common factors, not components, and factor loadings are typically lower than component loadings. Thus, because different methods of factor extraction were used, perfect fit would not be expected even if the correlation matrices in the normative and waiting list samples were identical. In the final model, this source of difference was eliminated. To estimate an upper limit of fit in the present data, we conducted an EMLFA in the waiting list sample itself, extracting five orthogonal factors. All 120 fixed parameters in the CFA were then set at those EMLFA values, and we used the CFA program to estimate the 30 remaining values and generate goodness-of-fit indices. This *EMLFA-based model* (Model 4) represents the best possible maximum likelihood fit of any five-factor model to these data.

### Results and Discussion

An examination of the maximum likelihood estimated factor loadings showed that the hypothesized NEO-PI-R structure was clearly recovered. At least 29 of the 30 facets showed loadings of .40 or higher on the hypothesized factor in each of the eight replication models (2a–3d); all 30 facets had loadings exceeding .40 in the most complete models (2d and 3d).

Table 1 presents goodness-of-fit indices for all 10 models. A variety of indices are presented for the benefit of interested readers; details on the calculation and interpretation of these indices are presented in Church and Burke ( 1994). In the present article, attention is focused on the chi-square values and on two widely used relative fit indices: the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) and the Normed Fit Index (NFI; Bentler & Bonett, 1980). For both of these indices, values of .90 are thought to represent good fit.

Chi-square values for all ten models are highly significant, meaning that none of the models provides a statistically significant fit in the waiting list data. Similarly, all the NFI indices are well below .90, and nine of the TLI values are less than .90—the exception being the orthogonal EMLFA-based model (4) derived from the waiting list sample itself. The orthogonal simple structure model (2a) shows the poorest fit. As gauged by TLI and NFI, fit is substantially improved by fixing secondary loadings at the normative values, and the more values specified, the better the fit-even when the loadings are quite small (less than ± .20 in Model 2d). It would appear from these latter analyses that the structure of the NEO-PI-R as seen in the normative sample can be replicated in the waiting list sample even in small details. These findings support the use of the full factor scoring matrix based on the normative data in computing factor scores for other data.

Oblique models as a group show the same pattern of fit indices as do the orthogonal models, with greater a priori specification of secondary loadings leading to better fit. The oblique simple structure model (3a) is clearly better than its orthogonal counterpart (2a; TLI = .56 vs. .52), but there is little advantage of the oblique complete model (3d) over the more parsimonious orthogonal complete model (2d; TLI = .83 vs. .83; parsimonious goodness-of-fit index = .67 vs. .69). An examination of factor intercorrelations explains this fact: In the simple structure model (3a), there are several substantial correlations: N with E (−.44) and with C (−.53), and 0 with E (.53) and A (.41). In the complete model (3d), these four correlations are reduced to −.32, .15, .14, and −. 11, respectively, and no other correlation exceeds .20 in absolute magnitude. It is quite clear in this analysis that the NEO-PI-R measures five distinct factors.[1]

Perhaps the most noteworthy result in Table 1 is the relatively poor fit of the EMLFA-based model (4), the best possible fit in these data for an orthogonal five-factor solution. In typical applications of CFA (e.g., Church & Burke, 1994), this result would lead to attempts to improve fit by respecifying the model. LISREL provides modification indices that identify specific problems in the residual matrix that can guide these respecifications. For example, in the present study the largest single problem was a residual correlation between NEO-PI-R facet scales E4: Activity and C4: Achievement Striving.

It would be possible to specify a correlated error term between these two scales, but the interpretation of such a term is unclear. *Correlated error* usually refers to a nonsubstantive source of variance. If Activity and Achievement Striving were, say, observer ratings, whereas all other variables were self-reports, it would make sense to control for this difference in method by introducing a correlated error term. But there are no obvious sources of correlated error among the NEO-PI-R facet scales in the present study.

## Study 2: Factor Replicability and the Number of Factors

An alternative conclusion to be drawn from the poor fit of Model 4 is that there are other, substantive, factors in the data beyond the first five-perhaps a small Industriousness factor defined by Activity and Achievement Striving. This is precisely the conclusion that tests of the number of factors (based on the chi-square statistic) yield in EMLFA analyses of the waiting list data, and the conclusion seems to be reinforced in the present instance by the fact that factor analysis of the waiting list data shows six eigenvalues greater than 1.0. When a six-factor EMLFA solution is examined, however, the chi-square test suggests that more factors are still needed. In fact, a statistically nonsignificant chi-square is not reached until 13 factors are extracted.

By almost any criterion other than the chi-square test, a 13-factor solution for the 30 NEO-PI-R facet scales would be considered overextraction. A scree test suggests 5 factors, as does Horn's ( 1965) parallel analysis method, the criterion recommended by Matthews and Oddy ( 1993; cf. Zwick & Velicer, 1986).[2] In the 13-factor solution 5 of the factors were defined

[1] After we completed these analyses, a version of LISREL 8 (Jöreskog & Sörbom, 1993) became available that includes Browne and Cudeck's ( 1993) measures of fit. Browne and Cudeck argued that a model shows a close fit if the root mean square error of approximation (RMSEA) is less than .05 and that "values up to .08 represent reasonable errors of approximation in the population" (Jöreskog & Sörbom, 1993, p. 124). Of the eight replication models, two meet this criterion: Model 2d (RMSEA = ,074) and Model 3d (RMSEA = ,075). By this standard, the fully specified normative model can be considered an acceptable fit.

[2] We conducted parallel analyses on five sets of random data with 30 variables and 229 "cases." The first six eigenvalues from these simulations were very similar, with means of 1.75, 1.63, 1.55, 1.49, 1.43, and I .40. In the waiting list sample, the eigenvalues were 6.04, 4.35, 3.35, 2.42, 1.65, and 1.04; only the first five eigenvalues exceeded their corresponding parallel random analysis eigenvalue, suggesting that five, and only five, factors are nonrandom.

Table 1
*Overall Goodness-of- Fit Indices for NEO-PI-R Models*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Absolute indices | | | | Relative indices | | Parsimony indices | |
| Model | $\chi^2$ | *df* | $\chi/df$ | GFI | RMS | TLI | NFI | CFI | AGFI | PGFI |
| 1. Null model | 3,643.5 | 435 | 8.38 | .36 | .24 | — | — | — | .32 | .33 |
| Orthogonal models | | | | | | | | | | |
| 2a. Simple structure | 1,846.6 | 405 | 4.56 | .63 | .18 | .52 | .49 | .55 | .57 | .55 |
| 2b. Salient loadings (>.4) | 1,649.0 | 405 | 4.07 | .66 | .16 | .58 | .55 | .61 | .57 | .61 |
| 2c. Modest loadings (>.2) | 1,065.7 | 405 | 2.63 | .77 | .11 | .78 | .71 | .79 | .73 | .67 |
| 2d. Complete | 911.9 | 405 | 2.25 | .79 | .09 | .83 | .75 | .84 | .76 | .69 |
| Oblique models | | | | | | | | | | |
| 3a. Simple structure | 1,679.5 | 395 | 4.25 | .64 | .13 | .56 | .54 | .60 | .58 | .55 |
| 3b. Salient loadings (>.4) | 1,561.9 | 395 | 3.95 | .67 | .15 | .60 | .57 | .64 | .62 | .57 |
| 3c. Modest loadings (>.2) | 1,016.8 | 395 | 2.57 | .78 | .10 | .79 | .72 | .81 | .74 | .66 |
| 3d. Complete | 904.2 | 395 | 2.29 | .79 | .09 | .83 | .75 | .84 | .75 | .67 |
| 4. EMLFA-based model | 681.5 | 435 | 1.57 | .84 | .04 | .92 | .81 | .92 | .82 | .79 |

*Note.* N = 229. Models 2b–2d and 3b–3d have fixed values based on a varimax-rotated principal-components analysis of data from the normative sample; Model 4 is based on a varimax-rotated maximum likelihood factor analysis of data from the waiting list sample. NEO-PI-R = Revised NEO Personality Inventory; GFI = goodness-of-fit index; RMS = root mean square; TLI = Tucker-Lewis Index; NFI = normed fit index; CFI = normed noncentrality fit index; AGFI = adjusted goodness-of-fit index; PGFI = parsimonious goodness-of-fit index; EMLFA = exploratory maximum likelihood factor analysis.

by only a single facet, and the last had no loading greater than ±.40. *By* definition, *common factors* are defined by at least two variables; are these small factors nevertheless to be considered in some sense meaningful?

Purely statistical criteria are not likely to yield a clear answer, but empirical studies ofreplicability may. As Thompson (1994) argued, “replicability analyses are attempts to look at data from perspectives intimately associated with the sine qua non of science-finding noteworthy effects that replicate under stated conditions” (p. 170). Similar logic lay behind Everett’s (1983) proposal that replicability can provide guidance in the problem of the number of factors. Everett proposed that the only meaningful factors are replicable factors and suggested that researchers examine factor comparability across subsamples of their data sets. Factor comparabilities are determined by correlating two sets of factor scores based on factor scoring matrices derived from two independent subsamples. Everett argued that the correct number of factors is indicated by the solution in which all the factors have comparabilities greater than .90. In recent years a number of researchers (Lanning, 1994; Matthews & Stanton, 1994; McCrae & Costa, *1987)* have adopted this approach.

Matthews and Oddy (1993) raised some objections to factor replicability as a criterion for the number of factors because the technique is sensitive to the size of loadings as well as to the size of the sample; in particular, when small samples are used, the number of factors tends to be underestimated. When adequate sample sizes are used, however, it is hard to understand how replicability would not be seen as a necessary condition for determining the number of factors in an exploratory analysis. There is no scientific utility in discovering the correct number of factors if we cannot reliably identify the factors because they fail to replicate from sample to sample. In this study we examined replicability of NEO-PI-R factors across and within the waiting list and normative samples.

### *Results and Discussion*

Table 2 reports results of a comparability analysis in the NEO-PI-R normative sample data that used factor scoring matrices derived from the normative sample (N = 1,000) and from the waiting list sample (N = 229). These sample sizes appear to be ample to detect replicable factors (Matthews & Oddy, 1993). We conducted two sets of analyses, the first using EMLFA and the second using principal-components analysis to extract factors. For example, the first line of Table 2 shows comparabilities for the 13-factor maximum likelihood solution: Thirteen factors were extracted separately in the two samples and, after varimax rotation, factor scoring coefficients were computed. These scoring coefficients were applied to the normative data to generate two sets of factor scores, which were then correlated. The highest value (.96) in the 13 X 13 matrix of correlations determined the first factor match, the highest value (.93) among the remaining factor correlations determined the second match, and so on until all 13 factors were paired. This process was repeated with decreasing numbers of extracted factors and with principal-components analyses.

The results shown in Table 2 are clear. When exactly five factors are extracted, very high comparabilities are found, all exceeding .92. When more than five factors are extracted, unreplicable factors appear in all analyses. By Everett’s (1983) criterion, no more than five factors in the NEO-PI-R are replicable across relatively large and comparable samples. Because one set of factor scoring weights was obtained from the normative sample, these results can be substantively interpreted to mean that after varimax rotation, the first five factors in the waiting list sample measure N, E, O, A, and C. The hypothesized NEO-PI-R structure is closely replicated in these exploratory factor analyses.

The chi-square test that indicated the need for 13 factors ap pears to have been unduly sensitive to unreplicable fluctuations in the correlation matrix. Perhaps that means there are real, if subtle,

Table 2
*Comparabilities for Varimax-Rotated Factors in the Norrnaiive Sample Using Factor Scoring Matrices Derived From the Normative and Waiting List Samples*

| | Factor comparabihties after varimax rotation | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Factors rotated | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th | 11th | 12th | 13th |
| Maximum likelihood | | | | | | | | | | | | | |
| 13 | .96 | .93 | .90 | .86 | .81 | .74 | .65 | .64 | .48 | .34 | .29 | .22 | .19 |
| 12 | .96 | .96 | .93 | .92 | .91 | .90 | .62 | .61 | .59 | .45 | .35 | .16 | |
| 11 | .97 | .96 | .95 | .95 | .78 | .78 | .66 | .51 | .43 | .21 | .08 | | |
| 10 | .98 | .93 | .90 | .78 | .77 | .77 | .70 | .28 | .23 | .oo | | | |
| 9 | .98 | .97 | .90 | .88 | .85 | .76 | .74 | .74 | .17 | | | | |
| 8 | .98 | .95 | .94 | .86 | .84 | .68 | .68 | .22 | | | | | |
| | .99 | .98 | .97 | .96 | .92 | .73 | .56 | | | | | | |
| 6 | .98 | .97 | .96 | .96 | .74 | .08 | | | | | | | |
| | .99 | .98 | .98 | .93 | .92 | | | | | | | | |
| Principal components | | | | | | | | | | | | | |
| 13 | .94 | .93 | .89 | .89 | .89 | .81 | .81 | .80 | .78 | .76 | 71 | .30 | .10 |
| 12 | .95 | .93 | .91 | .91 | .90 | .88 | .85 | .80 | .76 | .73 | .68 | .08 | |
| 11 | .96 | .93 | .92 | .89 | .83 | .83 | .82 | .76 | .76 | .72 | .44 | | |
| 10 | .95 | .92 | 91 | .79 | .78 | .71 | .70 | .68 | .53 | .45 | | | |
| 9 | .95 | .87 | .79 | .79 | .78 | .75 | .59 | .34 | .08 | | | | |
| 8 | .96 | .91 | .84 | .70 | .61 | .53 | .51 | .13 | | | | | |
| | .96 | .96 | .94 | .71 | .64 | .58 | .52 | | | | | | |
| | .97 | .96 | .92 | .91 | .71 | .20 | | | | | | | |
| 5 | .99 | .97 | .97 | .97 | .96 | | | | | | | | |

**Note. N** = 1,000.

differences between the two samples that result in different structures after the first five factors. The waiting list sample completed the test at a different time, was somewhat better educated than the normative sample, had a different proportion of men and women, and so on. Perhaps these sample differences lead to additional factors that would be replicable if truly comparable samples were considered.

One way to test that interpretation is by randomly subdividing the larger, normative sample into two subsamples. Because the division is random, the two groups can be presumed to be comparable in all respects. EMLFA chi-square statistics suggested the need for 14 factors in the first subsample, but computational problems made it impossible to examine a 14-factor solution in the second subsample. Table 3 shows the results of factor comparability analyses for 5- through 13-factor solutions. As in Table 2, only the 5-factor solution proved replicable.

An alternative approach to the number of factors is Velicer's ( 1976) minimum average partial (MAP) criterion, recommended by Lanning ( 1994; cf. Zwick & Velicer, 1986). This method examines the root mean square correlation among the variables after partialing the components extracted and selects the number of factors that minimizes this value. Extracting 1 through 14 components in the full normative sample leads to MAP values of. 189, .164, .136, .116, .111, .116, .124, .132, .139, .148, .159, .169, .181, and .193, respectively, suggesting a five-factor solution.

The comparability analyses, supported by the MAP analysis, suggest a serious problem for the chi-square test: The clear majority of the factors the test recommended are patently unreplicable. This is hardly a new finding. Montanelli ( 1974) found that the chi-square test was useful in simulated data that exactly fit the factor model but was "no use as a measure of goodness of fit for data that do not" (p. 555 ). Jöreskog ( 1974) himself noted that "the values of [ chi-square] should be interpreted very cautiously. In most empirical work many of the hypotheses may not be realistic" (p. 4).

The statistical logic behind the chi-square test is presumably sound; why, then, does it apparently yield the wrong conclusions? One possible answer is that the test is not robust, that some of the many assumptions that must be made in applying it to real data are untenable. In an extensive Monte Carlo evaluation, Hu et al. ( 1992) found that when some of the assumptions of multivariate normality were violated, maximum likelihood tests "for all practical purposes were completely useless at evaluating model adequacy at all sample sizes, because they almost always rejected the true model" (p. 358). Another possibility is that the problem lies not in the chi-square test but in the limitations of factor analysis: Perhaps there is indeed reliable residual variance, but too little, relative to error, to define replicable factors. Until statisticians and methodologists have resolved these problems it would seem unwise for one to put much reliance on the chi-square statistic itself when evaluating factor structures. It would seem that this caution should apply not only to the tests of significance (which have in fact long been disregarded) but also to those indices of fit-including the TLI and NFI—that are based on the chi-square. Until we know why nonrandom but nonreplicable residuals are left when personality scales are factored, questions must remain about indices based on those residuals.

## Study 3: An Alternative: Procrustes Rotation

Personality psychologists have been evaluating the replicability of factor structures for decades, using interpretive similarity

Table 3
*Comparabilities for Varimax-Rotated Maximum Likelihood Factors in the Normative Sample Using Factor Scoring Matrices Derived From Two Random Subsamples of the Normative Sample*

| Factors rotated | Factor comparabilities after varimax rotation | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th | 11th | 12th | 13th |
| 13 | .99 | .94 | .94 | .91 | .88 | .87 | .80 | .75 | .75 | .56 | .55 | .27 | .26 |
| 12 | .99 | .97 | .92 | .75 | .73 | .73 | .73 | .70 | .61 | .43 | .41 | .21 | |
| 11 | .99 | .97 | .87 | .86 | .76 | .72 | .72 | .61 | .57 | .55 | .05 | | |
| 10 | .99 | .95 | .94 | .90 | .89 | .86 | .67 | .62 | .56 | .20 | | | |
| 9 | .98 | .97 | .94 | .93 | .88 | .84 | .74 | .50 | .31 | | | | |
| 8 | .99 | .95 | .92 | .89 | .84 | .76 | .71 | .31 | | | | | |
| 7 | .99 | .98 | .96 | .95 | .75 | .73 | .72 | | | | | | |
| 6 | .99 | .99 | .98 | .96 | .93 | .45 | | | | | | | |
| 5 | .99 | .99 | .99 | .98 | .97 | | | | | | | | |

*Note.* *N* = 1,000 for comparabilities. N = 507 for the first subsample, 493 for the second subsample.

when different variables are factored and using various quantitative indices of factor similarity when the same variables are factored (Gorsuch, 1983; Guadagnoh & Velicer, 199 1). Coefficients of factor congruence (Wrigley & Neuhaus, 1955) are perhaps most commonly used (e.g., Bond, 1979). CFA is intended to offer an advance over these traditional methods. Among statisticians it has the appeal of mathematical elegance; for researchers it offers at least three advantages over the simple computation of factor congruences. First, it is a modeling procedure that allows factor structure to be guided by theory rather than by computational algorithms. Second, it offers statistical tests of fit that ought to improve on subjective evaluations guided by rules of thumb. Finally, CFA programs not only evaluate overall fit but also include modification indices that help pinpoint the sources of poor fit. Thus, CFA can be used in revising models and the theories from which they were derived. An attractive alternative to CFA would offer the same advantages while at the same time yielding results that square better with empirical generalizations. In Study 3 we outline such an alternative.

We are interested here in the question of whether the factor structure in a replication sample matches a hypothesized structure, where the hypotheses are derived either from theory or from previous empirical results. A common approach is to conduct an exploratory factor analysis in the replication data and evaluate its similarity to a known structure. Bond (1979), for example, used the eigenvalue-of-one rule to determine the number of factors in a study of personality traits in a Chinese sample; he then compared the five varimax-rotated principal components to those found in Norman's ( 1963) American sample.

Exploratory factor analyses provide straightforward tests of replicability. If two correlation matrices are identical or nearly so, independent exploratory analyses will yield the same results; when independent analyses give similar results they provide strong evidence of replicability. But exploratory analyses are not necessarily optimal for testing hypothesized models (Watkins, 1989). Even if the analysis is theoretically guided with respect to the number of factors extracted (an obvious requirement that is sometimes disregarded: e.g., Livneh & Livneh, 1989), factor rotation is not guided by theory. This can be a crucial issue, especially when the hypothesized structure is not simple: Small differences in the observed correlations can then yield large differences in the position of the axes, and the solutions may appear to be dramatically different. This is a familiar problem in analyses of circumplex models (McCrae & Costa, 1989b).

One promising alternative is least-squares targeted rotation, in which the data to be examined are rotated to maximum fit with a hypothesized target matrix. For example, Paunonen et al. ( 1992) examined the similarity in factor structures of scales from the Personality Research Form (PRF; Jackson, 1984) in Canada, Finland, Poland, and Germany. For each pair of samples, they began by rotating one set of factors to maximum similarity with a previously established structure (Skinner, Jackson, & Rampton, 1976); they then rotated the second set of factors to maximum similarity with the first and evaluated congruence of the factors. Using this procedure, they found five similar factors in each country.

Targeted rotations are called *Procrustes* rotations because they force the data, as much as possible, to conform to a predetermined structure (Digman, 1967). These techniques are frequently regarded with suspicion, largely on the strength of an article by Horn( 1967 ). Horn used oblique Procrustes rotation on random data and obtained factors that appeared to confirm a theoretical model of intelligence. He concluded that "random variables may be labeled arbitrarily and pushed into solutions that make quite 'good sense' " (p. 820). He noted, however, that such solutions tended to include very large correlations among the factors-as high as .90. Thus, convergence with the targeted variables was purchased at the expense of discrimination among factors.

Procrustes rotations are not necessarily misleading, however. Norman ( 1969 ) compared rotations of real data to "best" and "worst" targets and found that the data could not be forced into a good fit with the "worst" structure even by oblique Procrustes rotation. Even more conservative is orthogonal Procrustes rotation (Schönemann, 1966). In that method, factors are rotated to minimize the sums of squares of deviations from a target matrix, under the constraint of maintaining orthogonality. The technique realigns the position of the axes in the factor space without affecting their relative positions, just as multiple visual perspectives on a rigid object such as a table give different views without in the least changing the shape of the table. Orthogonal Procrustes rotation offers a powerful technique for hypothesis-guided rotation.

### *Quantifying Fit*

Of several measures of similarity between factors, the congruence coefficient proposed by Wrigley and Neuhaus ( 1955) is perhaps the most familiar. For each corresponding pair of factors it is calculated as the sum of the cross-products of two sets of column-normalized factor loadings. A value of .90 is typically considered necessary to define a matching factor (Barrett, 1986; Mulaik, 1972), although this is merely a rule of thumb.

One limitation of the coefficient of factor congruence is that it evaluates the factor as a whole. Particularly when there are many variables in an analysis, a high factor congruence coefficient may be seen even though a few critical variables do not load as intended. This is a case of special interest in cross-cultural research. It might well be true that the same broad personality factors can be recovered in many different cultures but that a few of the specific variables that define them differ from culture to culture. Self-reports of intelligence, for example, tend to load primarily on the 0 factor in American studies but on the C factor in Chinese studies ( McCrae, 1994b). Coefficients of factor congruence might not be sensitive to such a shift in a single variable, particularly if the total number of variables is large.

The informal method of simply identifying variables that do or do not load as expected can be problematic when arbitrary cutoff values are applied to a structure that is not simple. In one study, Variable X might load .39 on Factor I and .41 on Factor II; in a second study, it might load .41 on Factor I and .39 on Factor II. If we adopted the common convention of interpreting only loadings above .40, Variable X's loadings would appear to be unreplicable, although it is clear that there is no real difference between the two solutions.

To examine the replicability of factor loadings for individual variables, we propose that researchers calculate a variable congruence coefficient, computed with the same formula as the factor congruence coefficient, applied across the rows rather than down the columns of the factor pattern matrix (cf. Kaiser, Hunka, & Bianchini, 197 1). To the extent that variables show the same pattern of loadings across factors, they will tend to have high variable congruence coefficients, although no rule of thumb can yet be proposed to assess a good fit. Note that for this analysis the two sets of factors must be arranged in the same order, an automatic outcome of orthogonal Procrustes rotation.

For some purposes it may also be useful to calculate an overall index of congruence between two factor matrices. A total congruence coefficient can be calculated as the sum of the cross products of all corresponding elements in two matrix-normalized factor matrices.

Schönemann's ( 1966) Procrustes rotation minimizes the sum of squared differences between corresponding factor loadings. Because factor loadings can be interpreted as Cartesian coordinates of the variables in the factor space, this means that the Euclidean distance between corresponding variables is minimized; the two sets of variables are in effect superimposed as closely as possible. Whether Matrix A is rotated toward Matrix B, or Matrix B toward Matrix A, there is only one closest position; in consequence, variable and total congruence coefficients are identical. This property does not hold for factor congruence coefficients.

### *Statistical Assessment of Fit*

Orthogonal Procrustes rotation of one factor matrix to another maximizes the size of the total congruence coefficient in part because real factors will be optimally aligned and in part because the technique capitalizes on chance. Is there a way to determine the role of chance in a Procrustes fit? If so, it could be used as the basis for a statistical test of the significance of fit.

Paunonen et al. ( 1992) proposed a Monte Carlo solution, in which observed factor congruences are compared with the distribution of factor congruences obtained after Procrustes rotation of the data to random targets. If the fit of a real data set matched to a real target is greater than 95% of the fits of the same data matched to random targets, one can conclude with better than 95% confidence that the fit of the real data is not simply due to capitalization on chance. Paunonen et al. ( 1992) used this procedure to evaluate the conventional factor congruence coefficient. A similar procedure could also be used to establish critical values for the variable and total congruence coefficients proposed in this article. A variation on this method is used in this article.

### *A Method for Analyzing Factor Replicability*

In the remainder of this article we describe an approach to the analysis of NEO-PI-R data that should be useful in any evaluation of the replicability of factors from that instrument and might also be viewed as a model for evaluating other instruments. The method proposed here examines both exploratory and targeted rotations and has five steps:

1: *Specify the target structure.* A binary target of 1s and 0s can be used to specify hypothesized factor loadings; alternatively, the full factor pattern matrix from a previous study can be used as a target. The results of Study 1 suggest that all the factor loadings in the normative factor matrix of the NEO-PI-R are meaningful, so we recommend that the full 30 X 5 matrix be used as the target in analyses of that instrument. The analysis thus tests the hypothesis that the structure represented in the complete normative factor matrix is replicated with new data.

2: *Factor the data to be tested.* Extract the hypothesized number of factors or components and use varimax rotation to obtain exploratory factor loadings in the new sample. In analyses of the NEO-PI-R, researchers should extract live principal components, corresponding to the five components reported in normative results. Most applied factor analysts recognize that, with a reasonably large number of variables, components analyses yield results that are very similar to common factor analyses (Goldberg & Digman, 1994; Velicer & Jackson, 1990), and principal-components analysis has its own elegance of mathematical and computational simplicity.

3: *Perform a targeted rotation.* To examine the extent to which differences between the target and the varimax matrix are due solely to the orientation of the axes, use Schönemann's ( 1966) technique for orthogonal rotation to the target.

4: *Calculate congruences.* For the varimax rotation, calculate factor congruences and identify matching factors; the fit can be evaluated by standard conventions for interpreting factor replicability (e.g., Mulaik, 1972). If al1 five factors show reasonable matches with the target factors, it is possible to calculate variable and total congruence coefficients after reordering the varimax factors in the standard order. Factor, variable, and total congruence coefficients should next be calculated between the target matrix and the Procrustes-rotated replication matrix. The Appendix provides an SAS Interactive Matrix Language (IML; SAS Institute, 1989 ) program to perform the Procrustes rotation and compute congruence coefficients for NEO-PI-R data.

*5: Evaluate the significance of the Procrustes fit.* Researchers might perform their own Monte Carlo studies to generate distributions of random congruences, or they might use Paunonen's ( 1994) table or prediction formula. When the NEO-PI-R is being evaluated, an alternative is to use the critical values we report in the next section from a study of the structure of the NEO-PI-R in Chinese translation. Observed congruences are compared with the critical values to determine whether the fit can be regarded as chance.

## An Empirical Example

To examine the generalizability of the FFM to non-Indo-European cultures, the NEO-PI-R was translated into Chinese and administered to a sample of 352 college students in Hong Kong. Details on the process of translation and on the collection of data are presented elsewhere (McCrae, Costa, & Yik, 1996). From a cross-cultural perspective, it would be just as appropriate to ask how well the American data replicate the Chinese factor structure as vice versa, and thus the Chinese matrix might be used as the target. In this case, however, the American structure is known to be highly replicable in American samples, and considerable data on the construct validity of the American factors have been gathered: by contrast, little is known about the replicability or validity of the Chinese factors. The American factor matrix is therefore used as the target in this analysis. Recall, however, that variable and total congruence coefficients are invariant across the choice of target and in this sense are "culture-free" measures of factor structure replication.

An exploratory factor analysis of the 30 NEO-PI-R facet scales offered considerable support for the cross-cultural invariance of the FFM. Six factors had eigenvalues greater than 1 .0, but a scree test suggested just five factors. After varimax rotation, coefficients of congruence with the American normative factors ranged from .92 to .97, and 29 of the 30 facets loaded chiefly on the intended factor. Variable congruence coefficients ranged from .76 to .99; 26 of them were over .90. One A facet, Modesty, showed a larger loading on E (−.46) than on A (.32), and one 0 facet, Openness to Actions, had no loading greater than .29 (although that loading was on the intended factor).

Do these differences indicate a slightly different factor structure in Chinese samples, or are they the result of random perturbations of the location of the axes? A Procrustes rotation can help evaluate these possibilities. Following the steps outlined above, we rotated the five principal components to best fit the American normative structure. We then calculated factor, variable, and total congruences.

We used the Chinese data as the basis for a Monte Carlo study of the distribution of congruence coefficients following Procrustes rotation, adapted from the method recommended by ten Berge ( 1986). To generate random factor matrices with properties comparable to the original Chinese data, the 30 rows of factor loadings in the Chinese matrix were randomly permuted 1,000 times; a random half of the rows were then reflected. These simulations can be considered to represent possible FFMs, in which 30 scales define each of five factors but in which any combination of 6 scales per factor can occur. After each permutation, the new matrix was rotated to best fit the American normative target, and the three types of congruence coefficient were calculated. We then examined the distributions of these1,000 values, based on random data, for each factor, variable, and the total matrix. (A second Monte Carlo study, which we conducted using the same permutation procedure on a matrix of factors from purely random variables, yielded almost identical results.)

The means of the distributions of factor congruences based on Procrustes rotations of randomly permuted data ranged from .32 to .34 for the five factors. These low mean values clearly demonstrate that orthogonal Procrustes rotation cannot force random data into a spuriously close fit with a target. In fact, not 1 of the 5,000 factor congruence coefficients in the simulation reached .80, still less the .90 level that is traditionally considered evidence of factor replication.

The upper 95th and 99th percentiles of the distributions for the three types of congruence provide a basis for evaluating real data. The 95th percentiles for the five factor congruences ranged from .52 to .55; the 99th percentiles ranged from .59 to .65. Observed factor congruences exceeding .55 and .65 can therefore be considered non-chance with 95% and 99% confidence, respectively. Variable congruence coefficients were similar across the 30 variables: the mean of the 95th percentiles was .86; the mean of the 99th percentiles was .94. Finally, the 95th percentile for the total congruence coefficient was .42, and the 99th percentile was .46. These critical values can be used to assess the statistical significance of real data after Procrustes rotation.

Table 4 presents the Procrustes-rotated factor structure in the Chinese sample. In general, it is clear that the Chinese structure is very close to the normative target, supporting the cross-cultural replicability of the NEO-PI-R factor structure and the FFM. Every facet (including Modesty) has its highest loading on the intended factor, and the large secondary loadings for N2: Angry Hostility, E4: Activity, 03: Feelings, A3: Altruism, and Cl: Competence parallel large secondary loadings in the normative structure. The total congruence coefficient and each factor congruence coefficient is high and highly significant; 29 of the individual variables also showed high and significant variable congruences.

The single exception was 04: Actions, which was a rather weak definer of 0 in the Chinese data and showed an unexpectedly large negative loading on C. Strictly speaking, we cannot conclude from its nonsignificant variable congruence coefficient that this facet scale does not fit the hypothesized structure; we simply cannot assert that it does fit. In other Chinese sam-

Table 4
*Factor Loadings and Congruences for Factors in the Chinese NEO-PI-R Rotated to the Normative American Structure*

| | Factor | | | | | |
|---|---|---|---|---|---|---|
| NEO-PI-R facet | N | E | 0 | A | C | Variable congruence |
| Neuroticism (N) | | | | | | |
| N 1: Anxiety | **.84** | −.03 | −.05 | −.03 | −.07 | .99[b] |
| N2: Angry Hostility | **.64** | .11 | −.04 | **−.40** | −.11 | .98[b] |
| N3: Depression | **.77** | −.13 | −.06 | .03 | −.28 | .99[b] |
| N4: Self-Consciousness | **.67** | −.23 | −.14 | .03 | −.14 | .99[b] |
| N5: Impulsiveness | **.54** | .30 | .12 | −.23 | −.36 | .99[b] |
| N6: Vulnerability | **.74** | .04 | −.24 | .11 | −.34 | .96[b] |
| Extraversion (E) | | | | | | |
| E 1: Warmth | −.10 | **.74** | .11 | .33 | .11 | .99[b] |
| E2: Gregariousness | -.11 | **.70** | −.09 | .06 | −.08 | .97[b] |
| E3: Assertiveness | −.20 | **.57** | .11 | −.30 | .23 | .95[b] |
| E4: Activity | −.04 | **.59** | −.06 | −.18 | **.44** | .94[b] |
| E5: Excitement Seeking | −.01 | **.43** | .21 | −.39 | −.09 | .97[b] |
| E6: Positive Emotions | −.29 | **.61** | .21 | .16 | .08 | .93[a] |
| Openness (0) | | | | | | |
| 01: Fantasy | .15 | .02 | **.56** | −.15 | −.28 | .98[b] |
| 02: Aesthetics | .02 | .15 | **.67** | .25 | −.01 | .95[b] |
| 03: Feelings | **.49** | .27 | **.54** | .01 | .15 | .97[b] |
| 04: Actions | −.22 | .23 | .32 | −.15 | −.26 | .81 |
| 05: Ideas | −.12 | −.16 | **.60** | .01 | .33 | .93[a] |
| 06: Values | −.15 | .01 | **.43** | .06 | .05 | .88[a] |
| Agreeableness (A) | | | | | | |
| A 1: Trust | −.26 | .30 | .03 | **.62** | .13 | .96[b] |
| A2: Straightforwardness | −.06 | −.08 | .01 | **.72** | .13 | .97[b] |
| A3: Altruism | −.14 | **.40** | .18 | **.56** | .33 | .94[b] |
| A4: Compliance | −.09 | −.01 | −.07 | **.71** | −.09 | .98[b] |
| A5: Modesty | .33 | −.32 | −.17 | **.41** | −.17 | .88[a] |
| A6: Tender-Mindedness | .24 | .36 | .14 | **.56** | .05 | .95[b] |
| Conscientiousness (C) | | | | | | |
| C 1: Competence | **−.42** | .15 | .16 | −.20 | **.67** | .96[b] |
| C2: Order | −.08 | .02 | −.03 | .05 | **.73** | .97[b] |
| C3: Dutifulness | −.16 | −.02 | −.09 | .28 | **.69** | .99[b] |
| C4: Achievement Striving | −.06 | .16 | .04 | −.09 | **.79** | .98[b] |
| C5: Self-Discipline | −.33 | .10 | −.06 | .10 | **.75** | .99[b] |
| C6: Deliberation | −.27 | −.25 | .02 | .08 | **.67** | .97[b] |
| Factor/totalcongruence | .97[b] | .95[b] | .93[b] | .97[b] | .97[b] | .96[b] |

Note. N = 352. These are Procrustes-rotated principal components. Loadingsgreater than .40 in absolute magnitude are in boldface.

[a] Congruence higher than that of 95% of rotations from random data.

[b] Congruence higher than that of 99% of rotations from random data.

ples it might show a significant congruence. It is also possible, however, that the failure to reach a significant level may indicate the advisability of a revised translation or substitute items, or may suggest that need for variety in activities is not a central part of 0 in Chinese culture. Like modification indices in confirmatory factor analysis, variable congruences can be used to identify specific problem areas in fitting observed data to a target.

### *Two Illustrative Applications*

In one sense, the Chinese data do not illustrate the full power of Procrustes analyses, because the original varimax solution was itself very similar to the target matrix. The procedure is likely to be most useful when the fit of a factor solution to the hypothesized target is apparently poor. In those cases, Procrustes rotation can determine whether the lack of fit is merely a matter of the choice of axes or whether there are more substantial differences in structure.

Condo, Shimonaka, Nakazato, Ishihara, and Imuta ( 1993) reported results from a preliminary Japanese version of the NEO-PI-R. In their varimax-rotated factor analysis, N, O, and C factors were clear, showing factor congruence coefficients of .93–.94 with the corresponding American normative factors. E and A factors, however, were not easily identified. One of the remaining Japanese factors was defined by A1: Trust, E 1: Warmth, E6: Positive Emotions, A3: Altruism, E2: Gregariousness, and A6: Tender-Mindedness. The other was defined positively by E4: Activity, E3: Assertiveness, and E:5 Excitement Seeking, and negatively by A2: Straightforwardness, A4: Compliance, and A5: Modesty. These varimax factors could more easily be interpreted as the alternative interpersonal axes of Love and Dominance, respectively. Factor congruences with the American normative factors were low: .78 for Love with E and −.68 for Dominance with A.

Is this evidence of a failure to find cross-cultural replicability of the NEO-PI-R structure? No. After Procrustes rotation all five factor congruences with the normative target exceeded .92, with a total congruence coefficient of .94. Moreover, 27 of the 30 facets had statistically significant variable congruence coefficients. No variable congruence was lower than .80.

Figure 1 illustrates the relation between the Japanese and American factors and the way in which Procrustes rotation superimposes factors from two matrices. To construct Figure 1, we first plotted factor loadings for the six E and six A facets in the Japanese data against the varimax factors labeled Dominance and Love. The primary effect of the Procrustes analysis was to rotate these two factors about 35°; axes labeled E and A were therefore drawn at those positions in the figure. Next we plotted factor loadings for the E and A facets in the American normative structure with respect to the E and A axes in the figure, and corresponding Japanese and American facets were joined. An examination of the figure shows that all 12 facets occupy similar positions in the Japanese and American interpersonal planes and that either pair of axes (Dominance and Love or E and A) provides a reasonable basis for describing the data. This analysis clearly demonstrates that similar underlying factor structures can be concealed by varimax rotations and revealed by Procrustes rotations.

This article would not be complete without a final application of the proposed technique: How well do the waiting list data examined in Study 1 fit the hypothesized NEO-PI-R structure? Even varimax rotation produces an excellent replication of the normative structure, with congruence coefficients ranging from .94 to .99 (Costa & McCrae, 1995a; see also the factor comparabilities in Table 2). After Procrustes rotation of five principal components, total congruence with the normative target was .97, with individual factor congruences ranging from .95 to .98. All but 1 of the 30 facet scales showed a significant variable congruence coefficient. Data that yield at best a fair fit when CFA techniques are used show an excellent fit when evaluated by targeted rotation.

### *Sample Size and Significance Testing in Targeted Factor Rotation*

Sample size is an issue in interpreting CFA and Everett factor replicability procedures. How does it affect the proposed Monte Carlo evaluation of Procrustes rotated factors? In one sense it does not. Our simulation results began with random data, and the resulting critical values do not depend on sample size (Korth & Tucker, 1975; see also Paunonen, 1994). In another sense, of course, the validity of inferences about structure in a replication sample does depend on having a sufficiently large sample size, the stability of the observed structure being affected by the number of observations. Simulations by Gua-

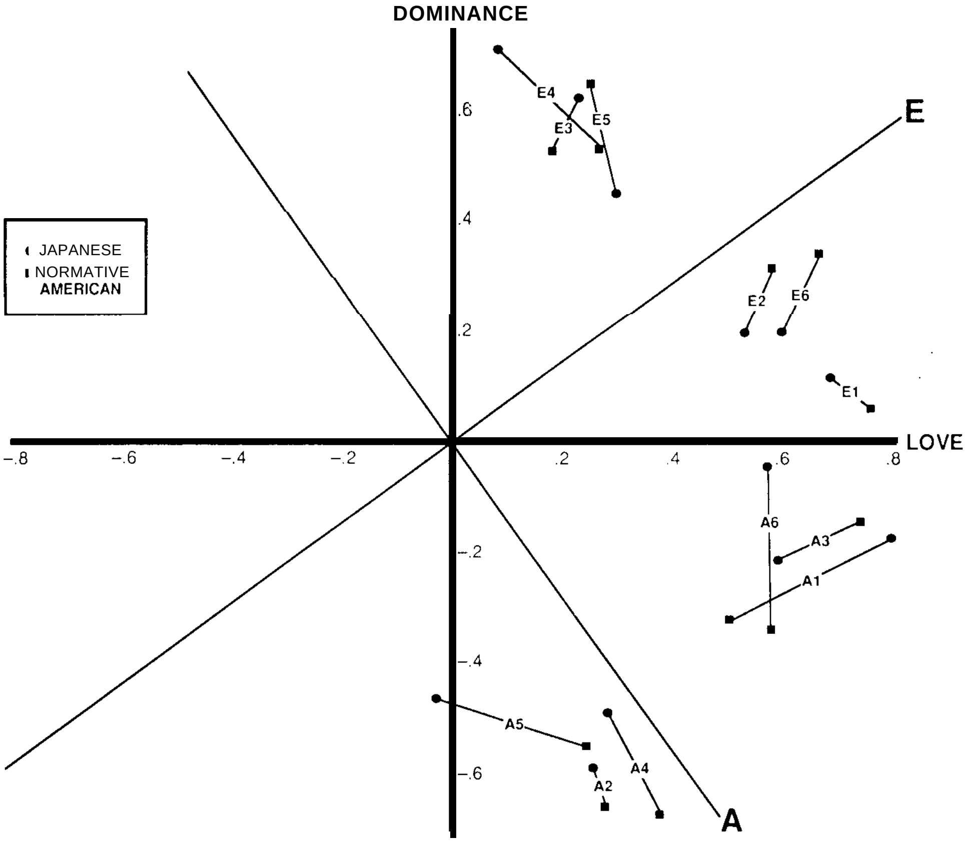

*Figure 1.* Factor plot of NEO-PI-R Extraversion (El to E6) and Agreeableness (Al to A6) facets in Japanese and American normative data. See Table 4 for facet scale labels.

dagnoli and Velicer ( 1988) suggest that a sample of 150 may be sufficient to show the NEO-PI-R structure. Increasing sample size is likely to give increasingly precise estimates of the population factor structure; if the true structure conforms to the hypothesized structure, then larger sample sizes will yield better fits. If the true structure departs from hypotheses, larger sample sizes will not necessarily lead to poorer fits, but they will give the researcher greater confidence that a poor fit is not simply due to sampling error.

The interpretation of "statistical significance" in the permutation method we have outlined requires some comment. Just as conventional tests for the significance of a Pearson correlation reveal only whether it is different from zero, and not whether it is meaningful, large, or within some theoretically expectable range, so the critical values we generated reveal only whether observed congruences exceed the values that Procrustes rotation might produce from random data. Significant values suggest that there is some non-zero similarity between the target and the rotated factors but do not tell us that the factors have beenreplicated.

It might well be argued that randomness is not the appropriate baseline for evaluating the replicability of factor structure. Researchers are not usually concerned that their data have no structure; they are interested in how well their structure matches a hypothesis. A translated version of the NEO-PI-R might well show non-chance resemblance to the American version but might also differ in important and replicable ways. In short, one might argue that, if CFA is unrealistically stringent in its criterion of fit, the present alternative is excessively lenient.

But statistical significance is not the only criterion by which hypotheses are evaluated, or necessarily the best (Cohen, 1994). Effect size also is important. In evaluating factor replicability, the size of the factor congruence coefficients is an index of the adequacy of fit, and the conventional rule of thumb of .90 is probably still meaningful. The value of the Monte Carlo simulation is primarily in showing that the meaningfulness of this rule is in no way compromised by the use of orthogonal Procrustes rotation, which virtually never produces such high values purely by chance. By the more stringent conventional criterion, the Chinese, Japanese, and waiting list data show excellent fit for all five factors in the NEO-PI-R.

## Conclusions

Confirmatory maximum likelihood factor analysis has a distinguished mathematical pedigree and is widely regarded by statisticians as the optimal way to evaluate a hypothesized factor structure. For many years the technique was so complex conceptually and so demanding computationally that only a handful of researchers used it. With increasing familiarity of the technique and the availability of convenient computer programs (e.g., Bentler, 1989; Jöreskog & Sörbom, 1993), it is likely that many more researchers will conduct CFA analyses in the future.

It is therefore essential to point out the dangers in an uncritical adoption and simplistic application of CFA techniques (cf. Breckler, 1990). In actual analyses of personality data from Borkenau and Ostendorf ( 1990 ) to Holden and Fekken (1994), structures that are known to be reliable showed poor fits when evaluated by CFA techniques. We believe this points to serious problems with CFA itself when used to examine personality structure.

If CFA has enjoyed an undeservedly favorable reputation, Procrustes rotations have suffered from the reverse. However, their recent use in conjunction with statistical tests of fit (Holden & Fekken, 1994; Paunonen et al., 1992; Stumpf, 1993) may point to a period of increasing acceptance. Prudent use of orthogonal Procrustes rotation in conjunction with Monte Carlo simulation techniques appears to lead to the acceptance of models that are replicable and the rejection of models that are not, and that is the ultimate test of the utility of a statistical procedure.

## References

Barrett, P. ( 1986). Factor comparison: An examination of three methods. *Personality and Individual Differences, 7,* 327–340.

Bentler, P. M. ( 1989). *EQS: A structural equations program manual.* Los Angeles: BMDP Statistical Software.

Bentler, P. M., & Bonett, D. G. ( 1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Block, J. ( 1961 ). *The Q-sort method in personality assessment and psychiatric research.* Springfield, IL: Charles C Thomas.

Bond, M. H. (1979). Dimensions of personality used in perceiving peers: Cross-cultural comparisons of Hong Kong, Japanese, American, and Filipino university students. *International Journal of Psychology, 14,* 47–56.

Borkenau, P., 8 Ostendorf, F. ( 1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11,* 515–524.

Breckler, S. J. ( 1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107, 260-273.*

Browne, M. W., & Cudeck, R. ( 1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology, 66,* 93– 114.

Cohen, J. (1994). The earth is round *(p* < .05). *American Psychologist, 49,* 997–1003.

Costa, P. T., Jr., & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences, 13, 653-665.*

Costa, P. T., Jr., & McCrae, R. R. (1992b). Reply to Eysenck. *Personality and Individual Differences, 13,* 861–865.

Costa, P. T., Jr., & McCrae, R. R. ( 1992c). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr.. & McCrae, R. R. (1992d). Trait psychology comes of age. In T. B. Sonderegger (Ed.), *Nebraska Symposium on Motivation.. Psychology and aging* (Vol. 39, pp. 169–204). Lincoln: University of Nebraska Press.

Costa, P. T., Jr., 8 McCrae, R. R. ( 1994). Stability and change in personality from adolescence through adulthood. In C. F. Halverson, G. A. Kohnstamm, & R. P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 139–150). Hillsdale, NJ: Erlbaum.

Costa, P. T., Jr., & McCrae, R. R. (1995a). Primary traits of Eysenck's P-E-N system: Three- and five-factor solutions. *Journal of Personality and Social Psychology, 69, 308-3 17.*

Costa, P. T., Jr., 8 McCrae, R. R. (1995b). Solid ground in the wetlands of personality: A reply to Block. *Psychological Bulletin, 117,* 216–*220.*

Costa, P. T., Jr., McCrae, R. R., 8 Dye, D. A. ( 1991 ). Facet scales for Agreeableness and Conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences, 12,* 887–898.

De Raad, B., Hendriks, A. A. J., & Hofstee, W. K. B. (1992). Towards a refined structure of personality traits. *European Journal of Personality, 6, 30*1–319.

Digman, J. M. ( 1967). The Procrustes class of factor-analytic transformations. *Multivariate Behavioral Research, 2, 89-94.*

Everett, J. E. ( 1983). Factor comparability as a means of determining the number of factors and their rotation. *Multivariate Behavioral Research. 18, 197-2*18.

Goldberg, L. R. ( 1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216– 1229.

Goldberg, L. R. ( 1993). The structure of phenotypic personality traits. *American Psychologist, 48, 26-34.*

Goldberg, L. R., 8 Digman, J. M. ( 1994). Revealing structure in the data: Principles ofexploratory factor analysis. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 216–242). New York: Springer.

Condo, Y., Shimonaka, Y., Nakazato, K., Ishihara, O., & Imuta, H. ( 1993, September). *Preliminary study for the standardization of the Japanese version of NEO-PI-R.* Paper presented at the 57th Meeting of the Japanese Psychological Association, Tokyo.

Gorsuch, R. L. ( 1983). *Factor analysis.* Hillsdale, NJ: Erlbaum.

Guadagnoli, E., & Velicer, W. F. ( 1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103,* 265–*275.*

Guadagnoli, E., & Velicer, W. F. ( 1991). A comparison of pattern matching indices. *Multivariate Behavioral Reseurch, 26, 323-343.*

Holden, R. R., 8 Fekken, G. C. ( 1994). The NEO Five-Factor Inventory in a Canadian context: Psychometric properties for a sample of university women. *Personality and Individual Differences, 17, 44*1–*444.*

Horn, J. L. ( 1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Horn, J. L. ( 1967). On subjectivity in factor analysis. *Educational and Psychological Measurement, 27, 8*11–820.

Hu, L., Bentler, P. M., 8 Kano, Y. ( 1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin. 112, 35*1–*362.*

Jackson, D. N. ( 1984). *Personality Research Form manual* (3rd. ed.). Port Huron, MI: Research Psychologists Press.

Jackson, D. N., & Chan, D. W. (1980). Maximum-likelihood estimation in common factor analysis: A cautionary note. *Psychological Bulletin. 88,* 502–508.

Joreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, O. H. Krantz, & R. 0. Suppes (Eds.), *Contemporary developments ın mathematical psychology* (Vol. 2, pp. 1–56). San Francisco: W. H. Freeman.

Joreskog, K. G., & Sörbom, D. (1988). *LISREL 7:A guide to the program and its application.* Chicago: SPSS.

Joreskog, K. G., & Sörbom, D. ( 1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Hillsdale, NJ: Erlbaum.

Kaiser, H. F., Hunka, S., & Bianchini, J. C. ( 1971 ). Relating factors between studies based upon different individuals. *Multivariate Behavioral Research, 6,* 409–422.

Korth, B., &Tucker, L. R. ( 1975). The distribution of chance congruence coefficients from simulated data. *Psychometrika, 40, 36*1–372.

Lanning, K. (1994). Dimensionahty of observer ratings on the Adult California Q-Set. *Journal of Personality and Social Psychology 67,* 151–160.

Livneh, H., & Livneh, C. (1989). The five-factor model of personality: Is evidence of its cross-measure validity premature? *Personality and Individual Differences, 10,* 75–80.

Marsh, H. W., Balla, J. R., & McDonald, R. P. ( 1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103, 39*1–410.

Matthews, G., & Oddy, K. ( 1993). Recovery of major personality dimensions from trait adjective data. *Personality and Individual Differences, 15,* 419–431.

Matthews, G., 8 Stanton, N. ( 1994). Item and scale factor analyses of the Occupational Personality Questionnaire. *Personality and Individual Differences, 16, 733-743.*

McCrae, R. R. ( 1989). Why I advocate the five-factor model: Joint analyses of the NEO-PI with other instruments. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 237-245 ). New York: Springer-Verlag.

McCrae, R. R. (1994a). The counterpoint of personality assessment: Self-reports and observer ratings. *Assessment, 1,* 159– 172.

McCrae, R. R. (1994b). Openness to Experience: Expanding the boundaries of Factor V. *European Journal of Personality, 8.25*1–272.

McCrae, R. R., & Costa, P. T., Jr. ( 1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology 52, 8*1–90.

McCrae, R. R., 8 Costa, P. T., Jr. (1989a). Rotation to maximize the construct validity of factors in the NEO Personality Inventory. *Multivariate Behavioral Reseurch, 24,* 107–124.

McCrae, R. R., & Costa, P. T., Jr. (1989b). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model, *Journal of Personality and Social Psychology, 56,* 586–595.

McCrae, R. R., Costa, P. T., Jr., & Busch, C. M. (1986). Evaluating comprehensiveness in personality systems: The California Q-Set and the five-factor model. *Journal of Personality 54,* 430–446.

McCrae, R. R., Costa, P. T., Jr., & Yik, M. S. M. ( 1996). Universal aspects of Chinese personality structure. In M. H. Bond (Ed.), *The handbook of Chinese psychology* (pp. 189–207). Hong Kong: Oxford University Press.

Montag, I.. & Levin, J. (1994). The five-factor personality model in applied settings. *European Journal of Personality, 8,* 1–11.

Montanelli, R. G., Jr. (1974). The goodness of fit of the maximum-likelihood estimation procedure in factor analysis. *Educational and Psychological Measurement, 34, 547-562.*

Mulaik, S. A. ( *1972). The foundations of factor analysis. New* York: McGraw-Hill.

Norman, W. 7 ( 1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology 66,* 574–583.

Norman, W. T. (1969). "To see oursels as ithers see us!": Relations among self-perceptions, peer-perceptions, and expected peer-perceptions of personality attributes. *Multivariate Behavioral Research, 4,* 417–443.

Ostendorf, F. ( 1990). *Sprache und Persönlichkeitsstruktur: Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit [ Language and persona&v structure: Toward the validation of the five-factor model of person&v].* Regensburg, Germany: S. Roderer Verlag.

Panter, A. T., Tanaka, J. S., & Hoyle, R. H. (1994). Structural models for multimode designs in personality and temperament research. In C. F. Halverson, G. A. Kohnstamm, 8 R. P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 111– 138). Hillsdale, NJ: Erlbaum.

Parker, J. D. A., Bagby, R. M., 8 Summerfeldt, L. J. (1993). Confirmatory factor analysis of the Revised NEO Personality Inventory. *Personality and Individual Differences, 15, 463-466.*

Paunonen, S. V.( 1994). On *chance and factor congruence following orthogonal Procrustes rotation.* Manuscript submitted for publication.

Paunonen, S. V., Jackson, D. N., Trzebinski, J., 8 Forsterling, F. ( 1992). Personality structure across cultures: A multimethod evaluation. *Journal of Personality and Social Psychology, 62,* 447–456.

Piedmont, R. L., 8 Weinstein, H. P. ( 1993). A psychometric evaluation of the new NEO-PI-R facet scales for Agreeableness and Conscientiousness. *Journal of Personality Assessment, 60, 302-3* 18.

SAS Institute, Inc. ( 1989). *SAS/IML software: Usage and reference, Version 6* ( 1st ed.). Cary, NC: Author.

Schönemann, P. H. ( 1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika, 31,*1– 10.

Shock, N. W., Greulich, R. C., Andres, R., Arenberg, D., Costa, P. T., Jr., Lakatta, E. G., & Tobin, J. D. ( 1984). *Normal human aging: The Baltimore Longitudinal Study of Aging* ( NIH Publication No. 84–2450). Bethesda, MD: National Institutes of Health.

Skinner, H. A., Jackson, D. N., & Rampton, G. M. (1976). The Personality Research Form in a Canadian context: Does language make a difference? *Canadian Journal of Behavioral Science, 8,* 156– *168.*

Stumpf, H. ( 1993). The factor structure of the Personality Research Form: A cross-national evaluation. *Journal of Personality, 61,* 27–48.

ten Berge, J. M. ( 1986). Rotation to perfect congruence and the cross-validation of component weights across populations. *Multivariate Behavioral Research, 21,* 41–64.

Thompson, B. ( 1994). The pivotal role of replication in psychological research: Empirically evaluating the replicabihty of sample results. *Journal of Personality 62,* 157– 116.

Tucker, L. R., & Lewis, C. ( 1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Velicer, W. F.( 1976). Determining the number ofcomponents from the matrix of partial correlations. *Psychometrika, 41,* 321–327.

Velicer, W. F., 8 Jackson, D. N. ( 1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,*1–28.

Watkins, D. ( 1989). The role of confirmatory factor analysis in cross-cultural research. *International Journal of Psychology 24, 685-70* 1.

Wiggins, J. S. ( 1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology 37,* 395–412.

Wrigley, C. S., & Neuhaus, J. 0. ( 1955). The matching of two sets of factors. *American Psychologist, 10, 4* 18-4 19.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99, 432-442.*

*(Appendix follows on next page)*

# Appendix

## SAS/IML Program

This program performs an orthogonal Procrustes rotation of a 30 X 5 matrix of principal-component factor loadings (rotated or unrotated) for NEO-PI-R facet scales. The data are entered between the braces for the matrix LOADINGS, with commas at the end of each line. Data in the matrix NORM are taken from Costa and McCrae,1992d, Table 2. The program prints a 3 1 X 6 matrix in which the sixth column gives variable and total congruences and the 3 1 st row gives factor congruences. Significance levels (see text) are indicated by asterisks beside the variable and total congruence coefficients and below the factor congruence coefficients. An SPSS version of this program is available from Robert R.McCrae.

```
PROC IML;
  LOADINGS ={
    };
  NORM ={
    .81    .02   -.Ol   −.01   −.10,
    .63   −.03    .01   −.48   −.08,
    .80   −.10    .02   −.03   −.26,
    .73   −.18   −.09    .04   −.16,
    .49    .35    .02   −.21   −.32,
    .70   −.15   −.09    .04   −.38,
   −.12    .66    .18    .38    .13,
   −.18    .66    .04    .07   −.03,
   −.32    .44    .23   −.32    .32,
    .04    .54    .16   −.27    .42,
    .oo    .58    .11   −.38   −.06,
   −.04    .74    .19    .10    .10,
    .18    .18    .58   −.14   −.31,
    .14    .04    .73    .17    .14,
    .37    41     .50   −.01    .12,
   −.19    .22    .57    .04   −.04,
   −.15   −.01    .75   −.09    .16,
   −.13    .08    .49   −.07   −.15,
   −.35    .22    .15    .56    .03,
   −.03   −.15   -.11    .68    .24,
   −.06    .52   −.05    .55    .27,
   −.16   −.08   -.oo    .77    .01,
    .19   −.12   −.18    .59   −.08,
    .04    .27    .13    .62    .00,
   −.41    .17    .13    .03    .64,
   −.04    .06   −.19    .01    .70,
   −.20   −.04    .Ol    .29    .68,
   −.09    .23    .15   −.13    .74,
   −.33    .17   −.08    .06    .75,
   −.23   −.28   −.04    .22    .57};
  S = LOADINGS`*NORM;
  W =EIGVEC(S*S`);
  V =EIGVEC(S`*S);
  0 = W`*S*V;
  K =DIAG(SIGN(O));
  WW =W*K;
  T =WW*V`;
  PROCRUST = LOADINGS*T;
  LABELS = {‘N 1’,‘N2’,‘N3’,‘N4’,‘N5’,‘N6’, ‘E 1’,‘E2’,‘E3’,‘E4’,‘E5’,
            ‘E6’, ‘O1’, ‘O2’, ‘O3’, ‘O4’, ‘O5’, ‘O6’, ‘ A1 ’, ‘A2’, ‘A3’,
            ‘A4’, ‘A5’, ‘A6’, ‘ C l ’ , ‘C2’, ‘C3’, ‘C4’, ‘C5’, ‘C6’,
            ‘FACTCONG’ };
  NAMES ={‘N’,‘E’,‘O’, ‘A’, ‘C’,‘Cong’};
  ROWP = {‘p’};
  A =(VECDIAG(NORM`*NORM))##{.5};
  B= (VECDIAG (PROCRUST`*PROCRUST))##{.5};
  C =VECDIAG((NORM`*PROCRUST)/(A*B`));
  D =(VECDIAG(NORM*NORM`))##{.5};
  E =(VECDIAG(PROCRUST*PROCRUST`))##{.5};
  F =VECDIAG((NORM*PROCRUST`)/(D*E`));
  G =(SUM(NORM#PROCRUST))/((SSQ(NORM))#
     (SSQ(PROCRUST)))##{.5};
  PROCCONG =(PROCRUST‖F)//(C`‖G);
  P5 = F > .86; P1 = F > .94; Y5 = G > .42; Y1 = G > .46;
  P = CHAR((P5 +P1)//(Y5+Y1)); Z5= C >.55; Zl = C >.65;
     Z= CHAR((Z5 +Z1)`);
  CALL CHANGE (P,‘           2’,‘**’, 0);
  CALL CHANGE (P,‘           1’,‘*’, 0);
  CALL CHANGE (P,‘           0’,‘’, 0);
  CALL CHANGE (Z,‘           2’,‘**’, 0);
  CALL CHANGE (Z,‘           1 ’,‘*’, 0);
  CALL CHANGE (Z, ‘          0’, ‘ ’, 0);
  RESET LINESIZE= 78 NONAME;
  PRINT ‘Procrustes Rotation with Congruence Coefficients’;
  PRINT PROCCONG [COLNAME = NAMES] [ROWNAME=
    LABELS] [FORMAT =8.2]P;
  PRINT Z [ROWNAME= ROWP] [FORMAT =$CHAR8.2];
  PRINT
‘*Congruence higher than that of 95% of rotations from random data.
**Congruence higher than that of 99% of rotations from random data.‘;
```

Received September 29, 1994
Revision received June 7, 1995
Accepted June 9, 1995 ■